Appendix for Medicaid Coverage Accuracy in Electronic Health Records

Visit-level Analysis description of two-stage logistic regression model

To simultaneously investigate patient-, visit-, and clinic-level characteristics associated with agreement in visit-level analyses, we used a two-stage logistic regression model that estimates odds ratios of agreement that controls for agreement due to chance as described in Lipsitz et al. 2003. This two-stage modeling approach was selected because a false association between the odds of agreement and a covariate could arise due entirely to chance agreement. The two-stage logistic regression model removed chance agreement in two stages. The first stage consisted of separate standard logistic regressions for each data source (EHR and Medicaid) with visit covered by Medicaid versus not Medicaid as the outcome. This stage estimated an offset which was utilized in the second stage to control for agreement due to chance. In the second stage, a single logistic regression model was performed where the outcome was agreement (agree versus not agree) between EHR and Medicaid data.

Let Y_{ijkr} equal 1 if the i-th visit for patient j in clinic k is denoted as covered by Medicaid from data source r and equal to 0 if it is denoted as not covered by Medicaid. Here, $i = 1, \ldots, m_j$, where m is the total number of visits for patient j; $j = 1, \ldots, n$ where n is the total sample size; $k = 1, \ldots, l$ where l is the total number of CHCs and $r = 1, 2$ denoting either the EHR data source or the Medicaid data source. Let Z_i be an indicator random variable which equals 1 if both data sources agree on Medicaid insurance coverage of the $i - th$ visit and 0 otherwise. In terms of Y_{ijk1} and Y_{ijk2}, we know that

$$Z_i = Y_{ijk1}Y_{ijk2} + (1 - Y_{ijk1})(1 - Y_{ijk2})$$

and thus its distribution is Bernoulli with probability $p_i = P(Z_i = 1|\boldsymbol{v}_{ijk}, \boldsymbol{x}_{ijk}, \boldsymbol{c}_{ijk})$ where $\boldsymbol{v}_{ijk}$ represent visit-level characteristics, $\boldsymbol{x}_{ijk}$ patient-level characteristics and $\boldsymbol{c}_{ijk}$ clinic-level characteristics.

The two-stage model to produce odds ratio of agreement accounting for chance agreement is as follows:

1. Perform separate logistic regressions of Y_{ijkr} on ($\boldsymbol{v}_{ijk}$, $\boldsymbol{x}_{ijk}$, $\boldsymbol{c}_{ijk}$) for $r = 1, 2$.
   For each visit, estimate the predicted probability of being covered by Medicaid to obtain $\hat{p}_{ijk1}$ and $\hat{p}_{ijk2}$. Use these predicted probabilities to estimate an offset term that will be used in the next stage that has the following form:
   $$\hat{\eta}_i = logit[\hat{p}_{ijk1}\hat{p}_{ijk2} + (1 - \hat{p}_{ijk1})(1 - \hat{p}_{ijk2})]$$

2. Perform a single logistic regression of Z_i on ($\boldsymbol{v}_{ijk}$, $\boldsymbol{x}_{ijk}$, $\boldsymbol{c}_{ijk}$) with $\hat{\eta}_i$ as an offset

$$logit(p_i) = \eta_i + \boldsymbol{v}_i^T\boldsymbol{\beta}_1 + \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \boldsymbol{c}_i^T\boldsymbol{\beta}_3$$

to obtain $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3)$ which represent the log odds ratio of agreement than would be expected under chance agreement.

Note that the estimated variance of $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3)$ reported by standard statistical software for logistic regression will not be correct because the offset is treated as known rather than estimated. Thus, we implemented a cluster bootstrap technique with 5,000 replicates to address this and well as to account for nesting of visits within patients and patients within clinics.

Patient-level Analysis description of two-stage logistic regression model

For patient-level analyses, we produced a 4-by-4 cross-tabulation of insurance cohort categories by data source and estimated agreement and kappa statistics. To simultaneously investigate patient- and clinic-level characteristics associated with agreement, we considered an extension of the two-stage logistic regression model described above to model multiple categories of insurance cohorts (continuously Medicaid, continuously not Medicaid, gained Medicaid, discontinuously Medicaid), instead of two categories (Medicaid and not Medicaid) as we did in visit-level analyses. The first stage consisted of performing separate multinomial logistic regression for each data source in order to obtain marginal probabilities of being assigned to a given insurance cohort to estimate the offset needed to adjust for chance agreement. The second stage implemented the offsets into the single logistic regression model with agreement between EHR and Medicaid data as the main outcome.

Let Y_{jkr} equal 1 if the j-th patient in clinic k is continously Medicaid, 2 if Continuously not Medicaid, 3 if Discontinuously Medicaid and 4 if Gained Medicaid. Let Z_j be an indicator random variable which equals 1 if both data sources agree on insurance cohort for the $j-th$ patient and 0 otherwise.

The two-stage model to produce odds ratio of agreement accounting for chance agreement is as follows:

1. Perform separate multinomial regressions of Y_{jkr} on ($\boldsymbol{x}_{jk}$, $\boldsymbol{c}_{jk}$) for $r = 1, 2$.

   For each patient, estimate the predicted probability of being in each insurance cohort to obtain ($\hat{p}^1_{jkr}$, $\hat{p}^2_{jkr}$,$\hat{p}^3_{jkr}$,$\hat{p}^4_{jkr}$) where $\hat{p}^g_{jkr}$ is the predicted probability of being in insurance cohort g ($g = 1, 2, 3, 4$) for data source r. Use these predicted probabilities to estimate an offset term that will be used in the next stage that has the following form:

$$\hat{\eta}_j = logit[\hat{p}^1_{jk1}\hat{p}^1_{jk2} + \hat{p}^2_{jk1}\hat{p}^2_{jk2} + \hat{p}^3_{jk1}\hat{p}^3_{jk2} + \hat{p}^4_{jk1}\hat{p}^4_{jk2}]$$

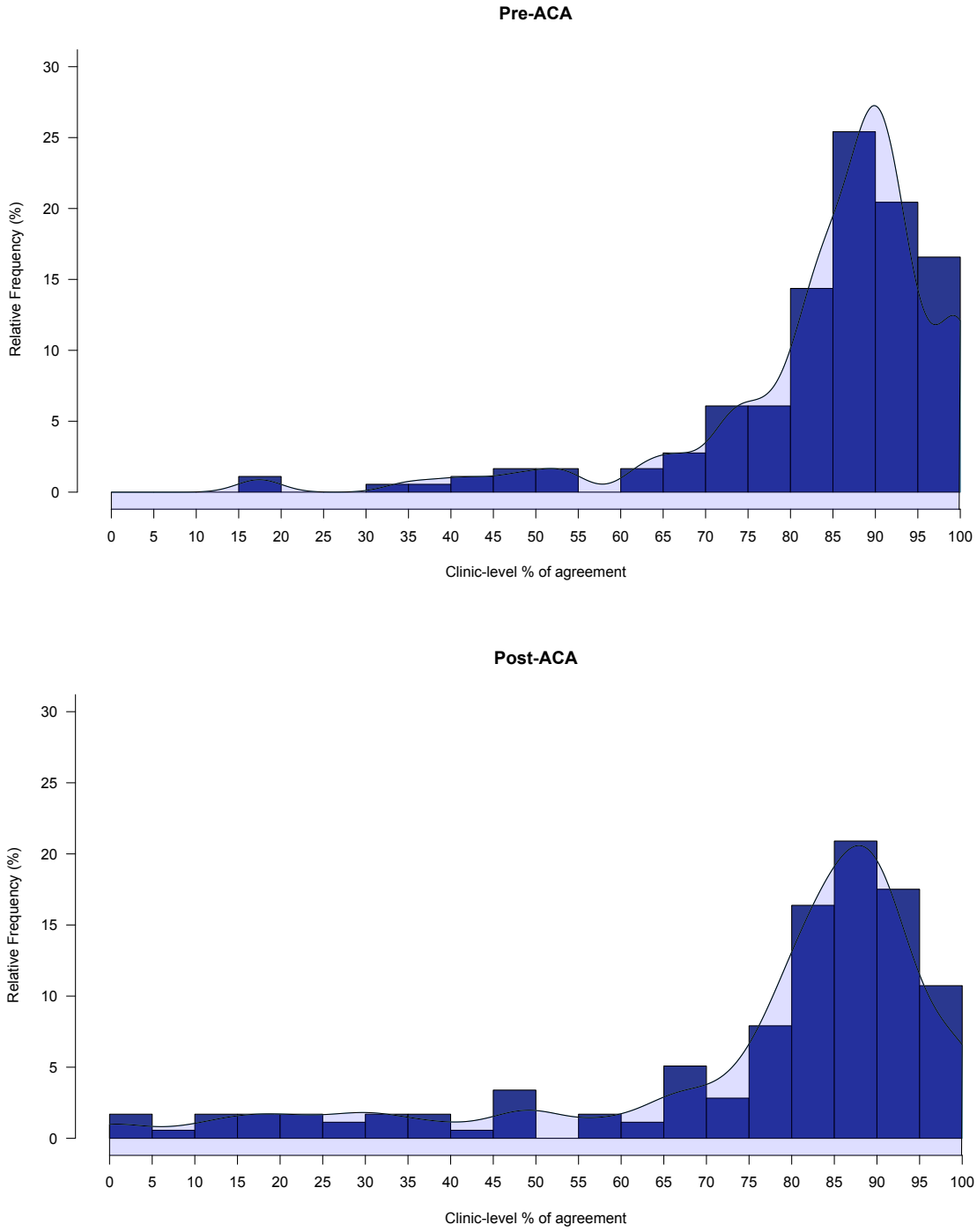
2. Perform a single logistic regression of Z_j on $(\boldsymbol{x}_{jk}, \boldsymbol{c}_{jk})$ with $\hat{\eta}_j$ as an offset

$$logit(p_j) = \eta_j + \boldsymbol{x}_j^T \boldsymbol{\beta}_1 + \boldsymbol{c}_j^T \boldsymbol{\beta}_2$$

to obtain $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ which represent the log odds ratio of agreement than would be expected under chance agreement.

Note that the estimated variance of $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ reported by standard statistical software for logistic regression will not be correct because the offset is treated as known rather than estimated. Thus, we implemented a cluster bootstrap technique with 5,000 replicates to address this and well as to account for nesting of patients within clinics.

Appendix Figure A.1. Histogram distribution of clinic-level Medicaid Coverage Agreement between EHR and Medicaid data, stratified by pre- and post-Affordable Care Act (ACA) periods



Note: Kernel Density Estimates of the distribution of clinic agreement is presented on top of the histogram bars. The kernel density estimator is a nonparametric method to estimate the probability distribution of agreement. Additionally, for this figure and for the analyses in the brief communication (e.g. two-stage logistic regression, boostrapping, etc.) we used the following R packages: *lubridate, broom, dplyr, forcats, multiwaycov, lmtest, boot, lme4.*

Appendix Figure A.2: Patients assigned to each insurance cohort using EHR and Medicaid data, N (column %)

Insurance cohorts based on EHR data	Insurance cohorts based on Medicaid enrollment data: Continuously Medicaid	Continuously not Medicaid	Discontinuous Medicaid	Gained Medicaid
Continuously Medicaid	**18,226 (78.8)**	2,977 (23.3)	4,051 (27.8)	220 (1.7)
Continuously not Medicaid	401 (1.7)	**5,582 (43.7)**	460 (3.2)	254 (1.9)
Discontinuous Medicaid	4,471 (19.3)	2,799 (21.9)	**6,060 (41.5)**	1,873 (14.1)
Gained Medicaid	20 (0.1)	1,402 (11.0)	4,020 (27.6)	**10,948 (82.3)**

Note: OCHIN health information network Epic© EHR data are referred to as EHR data and Oregon Medicaid enrollment data are referred to as Medicaid data. The shaded boxes denote when both EHR and Medicaid data classified a patient into the same insurance cohort. We defined insurance cohorts for both data sources the following way: 1) Continuously Medicaid: All visits in 2013 and 2014 covered by Medicaid; 2) Continuously Not Medicaid: All visits in 2013 and 2014 not covered by Medicaid (patients could have Medicare, private, VA/Military, worker's comp or no coverage); 3) Gained Medicaid: All visits in 2013 not covered by Medicaid and all visits in 2014 covered by Medicaid; and 4) Discontinuously Medicaid: Any combination of visit coverage that had discontinuous Medicaid coverage that did not follow the definition of the Gained Medicaid cohort. Overall, agreement was 64.0% with 95% confidence interval (CI)=63.6% – 64.4% and kappa was 0.504 with 95% CI=0.499 – 0.509.